Exploratory Data Analysis

Lecture 2

Corina Besliu

Technical University of Moldova

September 8, 2021

Outline

- Recapitulation
- Range
- Variance
- Standard Deviation
- Skewness
- Kurtosis

TECHNICAL UNIVERSITY OF MOLDOVA

Outline

- Recapitulation
- Range
- Variance
- Standard Deviation
- Skewness
- Kurtosis

Outline

- Recapitulation
- Range
- Variance
- Standard Deviation
- Skewness
- Kurtosis

Outline

- Recapitulation
- Range
- Variance
- Standard Deviation
- Skewness
- Kurtosis

TECHNICAL UNIVERSITY OF MOLDOVA

Outline

- Recapitulation
- Range
- Variance
- Standard Deviation
- Skewness
- Kurtosis

TECHNICAL UNIVERSITY OF MOLDOVA

Outline

- Recapitulation
- Range
- Variance
- Standard Deviation
- Skewness
- Kurtosis

TECHNICAL UNIVERSITY OF MOLDOVA

Lecture Outline

Recapitulation Slide

Descriptive Statistics

Last time we learned about the central location of the data.

- mean, median, mode

However, there are other characteristics that are important when analyzing data. That is measures of variability.

Variability is the degree of dispersion in the data

- range, variance, standard deviation, kurtosis, skewness

Recapitulation Slide

Descriptive Statistics

Last time we learned about the central location of the data.

- mean, median, mode

However, there are other characteristics that are important when analyzing data. That is measures of variability.

Variability is the degree of dispersion in the data

- range, variance, standard deviation, kurtosis, skewness

TECHNICAL UNIVERSITY OF MOLDOVA

Lecture Outline

Range Formula

Range = Largest observation - Smallest observation

- The advantage of the range is its simplicity.
- The disadvantage is also its simplicity.

Set 1: 10 12 12 12 15 16 17 100

Set 2: 10 30 40 50 60 70 80 90 100

Range = 90 for both sets

Yet the two sets of data are completely different.

We need other statistics that incorporate all the data and not just two observations!

Range Formula

Range = Largest observation - Smallest observation

- The advantage of the range is its simplicity.
- The disadvantage is also its simplicity.

Set 1: 10 12 12 12 15 16 17 100

Set 2: 10 30 40 50 60 70 80 90 100

Range = 90 for both sets

Yet the two sets of data are completely different.

We need other statistics that incorporate all the data and not just two observations!

Range Formula

Range = Largest observation - Smallest observation

- The advantage of the range is its simplicity.
- The disadvantage is also its simplicity.

Set 1: 10 12 12 12 15 16 17 100

Set 2: 10 30 40 50 60 70 80 90 100

Range = 90 for both sets

Yet the two sets of data are completely different.

We need other statistics that incorporate all the data and not just two observations!

Range Formula

Range = Largest observation - Smallest observation

- The advantage of the range is its simplicity.
- The disadvantage is also its simplicity.

Set 1: 10 12 12 12 15 16 17 100

Set 2: 10 30 40 50 60 70 80 90 100

Range = 90 for both sets

Yet the two sets of data are completely different.

We need other statistics that incorporate all the data and not just two observations!

Range Formula

Range = Largest observation - Smallest observation

- The advantage of the range is its simplicity.
- The disadvantage is also its simplicity.

Set 1: 10 12 12 12 15 16 17 100

Set 2: 10 30 40 50 60 70 80 90 100

Range = 90 for both sets

Yet the two sets of data are completely different.

We need other statistics that incorporate all the data and not just two observations!

Lecture Outline

Variance Definition & Formula

Variance refers to a statistical measurement of the spread between numbers in a data set. More specifically, variance measures how far each number in the set is from the mean.

- Population

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- Sample

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- Why do we square the deviations before averaging?
- Why for s^2 do we divide the sum of deviations by n-1 rather than n?

Variance Definition & Formula

Variance refers to a statistical measurement of the spread between numbers in a data set. More specifically, variance measures how far each number in the set is from the mean.

- Population

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- Sample

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- Why do we square the deviations before averaging?
- Why for s^2 do we divide the sum of deviations by n-1 rather than n?

Variance Definition & Formula

Variance refers to a statistical measurement of the spread between numbers in a data set. More specifically, variance measures how far each number in the set is from the mean.

- Population

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- Sample

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- Why do we square the deviations before averaging?
- Why for s^2 do we divide the sum of deviations by n-1 rather than n?

Variance Computation Example

Here we have the number of hours five students spent studying statistics last week:

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
8	$(8 - 7) = 1$	$(1)^2 = 1$
4	$(4 - 7) = -3$	$(-3)^2 = 9$
9	$(9 - 7) = 2$	$(2)^2 = 4$
11	$(11 - 7) = 4$	$(4)^2 = 16$
3	$(3 - 7) = -4$	$(-4)^2 = 16$
	$\sum_{i=1}^{5}(x_i - \bar{x}) = 0$	$\sum_{i=1}^{5}(x_i - \bar{x})^2 = 46$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1} = \frac{46}{5 - 1} = 11.5$$

Squared the deviations and thus squared measurement units make the interpretation of variance not very straightforward.

Variance Computation Example

Here we have the number of hours five students spent studying statistics last week:

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
8	$(8 - 7) = 1$	$(1)^2 = 1$
4	$(4 - 7) = -3$	$(-3)^2 = 9$
9	$(9 - 7) = 2$	$(2)^2 = 4$
11	$(11 - 7) = 4$	$(4)^2 = 16$
3	$(3 - 7) = -4$	$(-4)^2 = 16$
	$\sum_{i=1}^{5} (x_i - \bar{x}) = 0$	$\sum_{i=1}^{5} (x_i - \bar{x})^2 = 46$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} = \frac{46}{5 - 1} = 11.5$$

Squared the deviations and thus squared measurement units make the interpretation of variance not very straightforward.

Lecture Outline

Standard Deviation Definition & Formula

Standard Deviation is the square root of the Variance.

- Population

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

- Sample

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Standard deviation and Mean are arguably the most important statistics as they play a vital role in almost all statistical inference procedures.

Standard Deviation Definition & Formula

Standard Deviation is the square root of the Variance.

- Population

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

- Sample

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Standard deviation and Mean are arguably the most important statistics as they play a vital role in almost all statistical inference procedures.

Michael Jordan and Wilt Chamberlain

- Jordan and Chamberlain are basketball's most celebrated players.
- On average they both scored almost the same points per game
  - Mean - 30.12 for Jordan
  - Mean - 30.06 for Chamberlain
- However, when we consider standard deviation, Michael Jordan appears much more consistent
  - SD - Jordan: 4.76
  - SD - Chamberlain: 10.59

Michael Jordan and Wilt Chamberlain

- Jordan and Chamberlain are basketball's most celebrated players.
- On average they both scored almost the same points per game
  - Mean - 30.12 for Jordan
  - Mean - 30.06 for Chamberlain
- However, when we consider standard deviation, Michael Jordan appears much more consistent
  - SD - Jordan: 4.76
  - SD - Chamberlain: 10.59

Michael Jordan and Wilt Chamberlain

- Jordan and Chamberlain are basketball's most celebrated players.
- On average they both scored almost the same points per game
  - Mean - 30.12 for Jordan
  - Mean - 30.06 for Chamberlain
- However, when we consider standard deviation, Michael Jordan appears much more consistent
  - SD - Jordan: 4.76
  - SD - Chamberlain: 10.59

Michael Jordan and Wilt Chamberlain

- Jordan and Chamberlain are basketball's most celebrated players.
- On average they both scored almost the same points per game
  - Mean - 30.12 for Jordan
  - Mean - 30.06 for Chamberlain
- However, when we consider standard deviation, Michael Jordan appears much more consistent
  - SD - Jordan: 4.76
  - SD - Chamberlain: 10.59

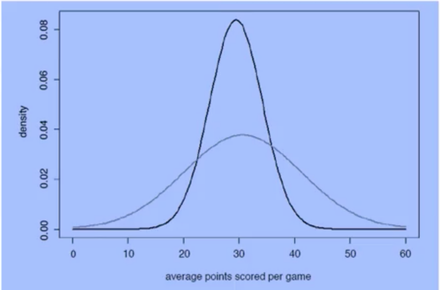


Figure 6.15 Normal distribution curves for Michael Jordan and Wilt Chamberlain

Reliability

- Measures of central tendency like Mean, Median and Mode can only paint a partial picture.
- Average statistics are incomplete without standard deviation/variance.
- Risk metrics are all about variance.

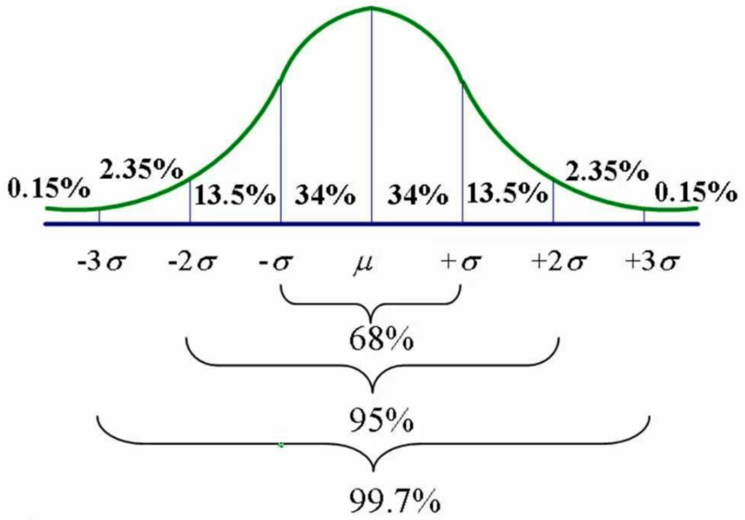
Interpretation

Knowing the mean and standard deviation allows the one to extract useful bits of information. The information depends on the shape of the histogram/distribution. If the distribution is bell shaped, we can use the Empirical Rule.

Empirical Rule

1. Approximately 68% of all observations fall within one standard deviation of the mean.
2. Approximately 95% of all observations fall within two standard deviations of the mean.
3. Approximately 99.7% of all observations fall within three standard deviations of the mean.

Empirical Rule



Chebysheffs Theorem

- If the distribution is not bell shaped, we can use the more general Chebysheffs Theorem.

Chebysheff's Theorem

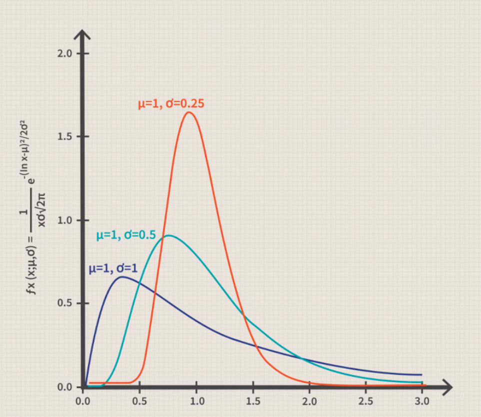
The proportion of observations in any sample or population that lie within k standard deviations of the mean is at least

$$1 - \frac{1}{k^2} \quad \text{for} \quad k > 1$$

Lecture Outline

Skewness Definition

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.



Skewness Coefficient and Properties

Skewness Coefficient

$$\tilde{\mu}_3 = \frac{\sum_i^N \left(X_i - \bar{X}\right)^3}{(N-1) * \sigma^3}$$

- Negative skew refers to a longer or fatter tail on the left side of the distribution, while positive skew refers to a longer or fatter tail on the right.
- The mean of positively skewed data will be greater than the median. In a distribution that is negatively skewed, the opposite is the case.
- If the data graphs symmetrically, the distribution has zero skewness, regardless of how long or fat the tails are.

Skewness Coefficient and Properties

Skewness Coefficient

$$\tilde{\mu}_3 = \frac{\sum_i^N \left(X_i - \bar{X}\right)^3}{(N-1) * \sigma^3}$$

- Negative skew refers to a longer or fatter tail on the left side of the distribution, while positive skew refers to a longer or fatter tail on the right.
- The mean of positively skewed data will be greater than the median. In a distribution that is negatively skewed, the opposite is the case.
- If the data graphs symmetrically, the distribution has zero skewness, regardless of how long or fat the tails are.

Skewness Coefficient and Properties

Skewness Coefficient

$$\tilde{\mu}_3 = \frac{\sum_i^N \left(X_i - \bar{X}\right)^3}{(N-1) * \sigma^3}$$

- Negative skew refers to a longer or fatter tail on the left side of the distribution, while positive skew refers to a longer or fatter tail on the right.
- The mean of positively skewed data will be greater than the median. In a distribution that is negatively skewed, the opposite is the case.
- If the data graphs symmetrically, the distribution has zero skewness, regardless of how long or fat the tails are.

Skewness Coefficient and Properties

Skewness Coefficient

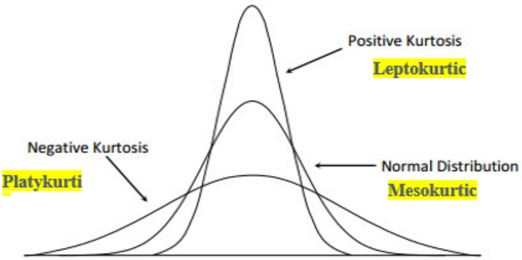
$$\tilde{\mu}_3 = \frac{\sum_i^N \left(X_i - \bar{X}\right)^3}{(N-1) * \sigma^3}$$

- Negative skew refers to a longer or fatter tail on the left side of the distribution, while positive skew refers to a longer or fatter tail on the right.
- The mean of positively skewed data will be greater than the median. In a distribution that is negatively skewed, the opposite is the case.
- If the data graphs symmetrically, the distribution has zero skewness, regardless of how long or fat the tails are.

Lecture Outline

Kurtosis

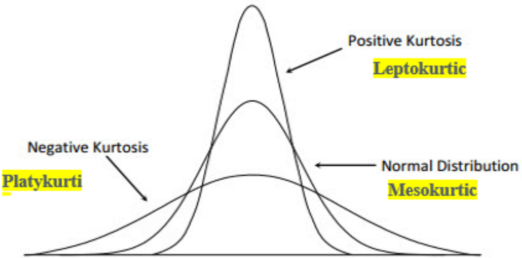
Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.



Fat tails indicate that observations are more extreme, i.e. we have more outliers.

Kurtosis

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.



Fat tails indicate that observations are more extreme, i.e. we have more outliers.

Kurtosis Formula and Properties

$$kurtosis = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)s^4}$$

- When a set of approximately normal data is graphed via a histogram, it shows a bell peak and most data within three standard deviations (plus or minus) of the mean.
- When high kurtosis is present, the tails extend farther than the three standard deviations of the normal bell-curved distribution.

e.g. For investors, high kurtosis of the return distribution implies the investor will experience occasional extreme returns (either positive or negative), – more extreme than the usual + or - three standard deviations from the mean that is predicted by the normal distribution of returns.

Kurtosis Formula and Properties

$$kurtosis = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)s^4}$$

- When a set of approximately normal data is graphed via a histogram, it shows a bell peak and most data within three standard deviations (plus or minus) of the mean.
- When high kurtosis is present, the tails extend farther than the three standard deviations of the normal bell-curved distribution.

e.g. For investors, high kurtosis of the return distribution implies the investor will experience occasional extreme returns (either positive or negative), – more extreme than the usual + or - three standard deviations from the mean that is predicted by the normal distribution of returns.

Kurtosis Formula and Properties

$$kurtosis = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)s^4}$$

- When a set of approximately normal data is graphed via a histogram, it shows a bell peak and most data within three standard deviations (plus or minus) of the mean.
- When high kurtosis is present, the tails extend farther than the three standard deviations of the normal bell-curved distribution.

e.g. For investors, high kurtosis of the return distribution implies the investor will experience occasional extreme returns (either positive or negative), – more extreme than the usual + or - three standard deviations from the mean that is predicted by the normal distribution of returns.

Kurtosis Formula and Properties

$$kurtosis = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)s^4}$$

- When a set of approximately normal data is graphed via a histogram, it shows a bell peak and most data within three standard deviations (plus or minus) of the mean.
- When high kurtosis is present, the tails extend farther than the three standard deviations of the normal bell-curved distribution.

e.g. For investors, high kurtosis of the return distribution implies the investor will experience occasional extreme returns (either positive or negative), – more extreme than the usual + or - three standard deviations from the mean that is predicted by the normal distribution of returns.

Kurtosis Formula and Properties

$$kurtosis = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)s^4}$$

- When a set of approximately normal data is graphed via a histogram, it shows a bell peak and most data within three standard deviations (plus or minus) of the mean.
- When high kurtosis is present, the tails extend farther than the three standard deviations of the normal bell-curved distribution.

e.g. For investors, high kurtosis of the return distribution implies the investor will experience occasional extreme returns (either positive or negative), – more extreme than the usual + or - three standard deviations from the mean that is predicted by the normal distribution of returns.

Lecture Outline

Let's get Started!

Access Google Colaboratory through your Gmail account